

A UNIVERSAL ERROR BOUND IN THE CLT FOR COUNTING MONOCHROMATIC EDGES IN UNIFORMLY COLORED GRAPHS

XIAO FANG *

National University of Singapore and Stanford University

Abstract

Let $\{G_n : n \geq 1\}$ be a sequence of simple graphs. Suppose G_n has m_n edges and each vertex of G_n is colored independently and uniformly at random with c_n colors. Recently, Bhattacharya, Diaconis and Mukherjee (2013) proved universal limit theorems for the number of monochromatic edges in G_n . Their proof was by the method of moments, and therefore was not able to produce rates of convergence. By a non-trivial application of Stein's method, we prove that there exists a universal error bound for their central limit theorem. The error bound depends only on m_n and c_n , regardless of the graph structure.

Key words and phrases: Stein's method; normal approximation; rate of convergence; monochromatic edges.

AMS 2010 Subject Classification: 05C15, 60F05

1 INTRODUCTION

Let $\{G_n : n \geq 1\}$ be a sequence of simple graphs, that is, graphs that contain no loops and no multiple edges. Suppose G_n has m_n edges and each vertex of G_n is colored independently and uniformly at random with c_n colors. Let Y_n be the number of monochromatic edges in G_n . Using the coupling approach in Stein's method for Poisson approximation, Barbour, Holst and Janson (1992) (page 105, Theorem 5.G) proved that

$$d_{TV}(\mathcal{L}(Y_n), Poi(\frac{m_n}{c_n})) \leq \frac{\sqrt{8m_n}}{c_n} \quad (1.1)$$

where d_{TV} denotes the total variation distance and $Poi(\lambda)$ denotes the Poisson distribution with mean λ . The bound (1.1) implies that if $c_n \rightarrow \infty$ and $m_n/c_n \rightarrow \lambda > 0$, the distribution of Y_n converges to the Poisson distribution with mean λ . Recently, Bhattacharya, Diaconis and Mukherjee (2013) reproved this Poisson limit theorem by the method of moments. By the same method, they also showed that in the case $c_n \rightarrow \infty$ and $m_n/c_n \rightarrow \infty$, the distribution of W_n , after proper standardization, converges to the standard normal distribution. These limit theorems

*Department of Statistics and Applied Probability, Block S16, Level 7, 6 Science Drive 2, Faculty of Science, National University of Singapore, Singapore 117546

were called *universal* limit theorems because they do not require any assumption on the graph structure. For applications of this and related problems, we refer to Bhattacharya, Diaconis and Mukherjee (2013) and the references therein.

In this note, we prove the following result.

Theorem 1.1. *Let Y be the number of monochromatic edges in a simple graph with m edges where each vertex is colored independently and uniformly at random with c colors. Let*

$$W = \frac{(Y - \frac{m}{c})}{\sqrt{\frac{m}{c}(1 - \frac{1}{c})}}.$$

We have

$$d_W(\mathcal{L}(W), N(0, 1)) \leq \frac{3}{2} \sqrt{\frac{c}{m}} + \frac{5\sqrt{2}}{\sqrt{c}} + \frac{1}{\sqrt{\pi}} \frac{2^{7/4}}{m^{1/4}}. \quad (1.2)$$

where d_W denotes the Wasserstein distance and $N(0, 1)$ denotes the standard normal distribution.

The bound (1.2) provides a universal error bound for the central limit theorem for W_n as $c_n \rightarrow \infty$ and $m_n/c_n \rightarrow \infty$. A corollary for fixed c_n is also obtained in Remark 2.5. The bound (1.2) is obtained by a non-trivial application of Stein's method for normal approximation. Stein's method was introduced by Stein (1972) for normal approximation. Stein's method for Poisson approximation was first studied by Chen (1975) and popularized by Arratia, Goldstein and Gordon (1990). We refer to Barbour and Chen (2005) for an introduction to Stein's method. Stein's method has been widely used to prove limit theorems with error bounds in graph theory. For example, Arratia, Goldstein and Gordon (1990) and Chatterjee, Diaconis and Meckes (2005) used Stein's method to prove Poisson limit theorems for monochromatic cliques in a uniformly colored complete graph. Cerquetti and Fortini (2006) considered more general monochromatic subgraphs counts when the distribution of colors was exchangeable. Janson and Nowicki (1991) studied the asymptotic distribution of the number of copies of a given graph in various random graph models. All of these results are obtained by exploiting the local dependence structure within random variables. In addition to the local dependence structure, we also exploit the uncorrelatedness within W_n . This technique of exploiting the uncorrelatedness within random variables was also used in Fang and Röllin (2014) to obtain rates of convergence for the central limit theorem for subgraph counts in random graphs.

In the next section, we give the proof of Theorem 1.1.

2 NORMAL APPROXIMATION

Let $G = (V(G), E(G))$ be a simple undirected graph, where $V(G)$ is the vertex set and $E(G)$ is the edge set. Let $m = |E(G)|$ be the number of edges of G . We color each vertex of G independently and uniformly at random with $c \geq 2$ colors. Formally, label the vertices of G by $\{v_1, \dots, v_{|V(G)|}\}$ and denote the color of the vertex v_i by ξ_{v_i} . Label the edges of G by $\{1, \dots, m\}$. For each edge i , we denote by v_{i1}, v_{i2} the two vertices it connects, i.e., $i = (v_{i1}, v_{i2})$. Without loss of generality, assume $\deg(v_{i1}) \leq \deg(v_{i2})$ where $\deg(v)$ denotes the degree of vertex v . Using the

above notation, the standardized number of monochromatic edges can be expressed as

$$W = \sum_{i=1}^n X_i := \sum_{i=1}^m \frac{(\mathbb{I}(\xi_{v_{i1}} = \xi_{v_{i2}}) - \frac{1}{c})}{\sqrt{\frac{m}{c}(1 - \frac{1}{c})}}. \quad (2.1)$$

Observing that X_i and X_j are uncorrelated if $i \neq j$, we have $\mathbb{E}W = 0$, $\text{Var}(W) = 1$.

We will need the following lemmas in the proof of Theorem 1.1.

Lemma 2.1. *We have the following bounds on the moments of X_i :*

$$\mathbb{E}|X_i| \leq \frac{2}{\sqrt{mc}}; \quad \mathbb{E}X_i^2 = \frac{1}{m}; \quad \mathbb{E}|X_i|^3 \leq \frac{\sqrt{c}}{m^{3/2}}. \quad (2.2)$$

Proof. The proof is elementary and therefore omitted. \square

Lemma 2.2 (Page 37 of Barbour, Holst and Janson (1992)). *For each edge $i = (v_{i1}, v_{i2})$, define $d_i = \deg(v_{i1}) \wedge \deg(v_{i2})$. We have*

$$\sum_{i=1}^m d_i \leq \sqrt{2}m^{3/2}. \quad (2.3)$$

Lemma 2.3 (Lemma 3.2 of Bhattacharya, Diaconis and Mukherjee (2013)). *The number of triangles, denoted by $\#(\Delta)$, in G is bounded by $\sqrt{2}m^{2/3}/3$.*

The following proposition is the key ingredient in proving Theorem 1.1.

Proposition 2.4. *For any function f with bounded first and second derivatives, we have with W defined in (2.1),*

$$|\mathbb{E}f'(W) - \mathbb{E}Wf(W)| \leq \|f''\| \left(\frac{3}{2} \sqrt{\frac{c}{m}} + \frac{5\sqrt{2}}{\sqrt{c}} \right) + \|f'\| \frac{2 \cdot 2^{1/4}}{m^{1/4}} \quad (2.4)$$

where $\|g\| := \sup_x |g(x)|$ for any function g .

Proof. For each edge $i = (v_{i1}, v_{i2})$ with $\deg(v_{i1}) \leq \deg(v_{i2})$, define the neighborhood $N_i \subset \{1, \dots, m\}$ to be consisted of all the edges connects to v_{i1} . Let

$$D_i = \sum_{j \in N_i} X_j, \quad W_i = W - D_i.$$

Since the color of v_{i1} is independent of W_i , we have X_i is independent of W_i . Therefore, by $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1/m$, the Taylor expansion and adding and subtracting corresponding terms, we have

$$\begin{aligned} \mathbb{E}f'(W) - \mathbb{E}Wf(W) &= \sum_{i=1}^m \mathbb{E}X_i^2 \mathbb{E}f'(W) - \sum_{i=1}^m \mathbb{E}X_i [f(W) - f(W_i)] \\ &= \sum_{i=1}^m \mathbb{E}X_i^2 \mathbb{E}f'(W) - \sum_{i=1}^m \mathbb{E}X_i D_i f'(W - UD_i) \\ &= \sum_{i=1}^m \mathbb{E}X_i^2 \mathbb{E}f'(W) - \sum_{i=1}^m \mathbb{E}X_i^2 f'(W - UD_i) - \sum_{i=1}^m \mathbb{E}X_i (D_i - X_i) f'(W - UD_i) \\ &=: R_1 - R_2 - R_3 - R_4 \end{aligned} \quad (2.5)$$

where U is an independent random variable distributed uniformly in $[0, 1]$ and

$$\begin{aligned} R_1 &= \sum_{i=1}^m \mathbb{E} X_i^2 \mathbb{E} f'(W) - \sum_{i=1}^m \mathbb{E} X_i^2 \mathbb{E} f'(W_i), \\ R_2 &= \sum_{i=1}^m \mathbb{E} X_i^2 [f'(W - U D_i) - f'(W_i)], \\ R_3 &= \sum_{i=1}^m \mathbb{E} X_i (D_i - X_i) [f'(W - U D_i) - f'(W)], \\ R_4 &= \mathbb{E} f'(W) \sum_{i=1}^m X_i (D_i - X_i). \end{aligned}$$

First of all, by the Taylor expansion,

$$|R_1| \leq \|f''\| \sum_{i=1}^m \mathbb{E} X_i^2 \mathbb{E} |D_i| \leq \|f''\| \left(\sum_{i=1}^m \mathbb{E} |X_i|^3 + \sum_{i=1}^m \sum_{j \in N_i \setminus \{i\}} \mathbb{E} |X_i|^2 \mathbb{E} |X_j| \right).$$

By (2.2) and (2.3),

$$\sum_{i=1}^m \mathbb{E} |X_i|^3 \leq \sqrt{\frac{c}{m}}, \quad \sum_{i=1}^m \sum_{j \in N_i \setminus \{i\}} \mathbb{E} |X_i|^2 \mathbb{E} |X_j| \leq \frac{1}{m} \frac{2}{\sqrt{mc}} \sum_{i=1}^m d_i \leq \frac{2\sqrt{2}}{\sqrt{c}}.$$

Therefore,

$$|R_1| \leq \|f''\| \left(\sqrt{\frac{c}{m}} + \frac{2\sqrt{2}}{\sqrt{c}} \right).$$

By the same argument and the fact that $\{X_j : j \in N_i\}$ are jointly independent,

$$|R_2| \leq \frac{1}{2} \|f''\| \left(\sqrt{\frac{c}{m}} + \frac{2\sqrt{2}}{\sqrt{c}} \right).$$

For R_3 , by the Taylor expansion,

$$|R_3| \leq \frac{1}{2} \|f''\| \sum_{i=1}^m \mathbb{E} |X_i| |D_i - X_i| |D_i| \leq \frac{1}{2} \|f''\| \sum_{i=1}^m \mathbb{E} X_i^2 |D_i - X_i| + \frac{1}{2} \|f''\| \sum_{i=1}^m \mathbb{E} |X_i| |D_i - X_i|^2.$$

Again by (2.2) and (2.3) and the fact that $\{X_j : j \in N_i\}$ are jointly independent,

$$\begin{aligned} \sum_{i=1}^m \mathbb{E} X_i^2 |D_i - X_i| &\leq \frac{1}{m} \frac{2}{\sqrt{mc}} \sum_{i=1}^m d_i \leq \frac{2\sqrt{2}}{\sqrt{c}}, \\ \sum_{i=1}^m \mathbb{E} |X_i| |D_i - X_i|^2 &\leq \frac{2}{\sqrt{mc}} \frac{1}{m} \sum_{i=1}^m d_i \leq \frac{2\sqrt{2}}{\sqrt{c}}. \end{aligned}$$

Therefore,

$$|R_3| \leq \|f''\| \frac{2\sqrt{2}}{\sqrt{c}}.$$

Finally we bound $|R_4|$. By the Cauchy-schwartz inequality, the fact that $\{X_j : j \in N_i\}$ are jointly independent and $\mathbb{E}X_i = 0$,

$$|R_4| \leq \|f'\| \sqrt{\text{Var}\left(\sum_{i=1}^m X_i(D_i - X_i)\right)} = \|f'\| \sqrt{\text{Var}\left(\sum_{i=1}^m \sum_{j \in N_i \setminus \{i\}} X_i X_j\right)}.$$

Observe that if $j \in N_i \setminus \{i\}$ and $l \in N_k \setminus \{k\}$, $\text{Cov}(X_i X_j, X_k X_l) = 0$ unless $\{i, j\} = \{k, l\}$ or $\{i, j, k, l\}$ forms a triangle. For the case $\{i, j\} = \{k, l\}$,

$$\text{Cov}(X_i X_j, X_i X_j) = \frac{1}{m^2}.$$

For the case $\{i, j, k, l\}$ forms a triangle, with distinct i, j, k ,

$$\text{Cov}(X_i X_j, X_j X_k) = \mathbb{E}X_i X_j^2 X_k \leq \frac{1}{m^2}$$

where the last inequality is by straightforward calculation. Therefore, by (2.3), Lemma 2.3, and observing that each triangle in G gives rise to 3 ordered pairs of (i, j) such that $j \in N_i \setminus \{i\}$, we have,

$$|R_4| \leq \|f'\| \sqrt{2 \sum_{i=1}^m \sum_{j \in N_i \setminus \{i\}} \frac{1}{m^2} + \frac{3 \times 2}{m^2} \#(\Delta)} \leq \|f'\| \frac{2 \cdot 2^{1/4}}{m^{1/4}}.$$

The bound (2.4) follows from (2.5) and the bounds on $|R_1| - |R_4|$. \square

Proof of Theorem 1.1. By the definition of Wasserstein distance, we have

$$d_W(\mathcal{L}(W), N(0, 1)) = \sup_{\|h'\| \leq 1} |\mathbb{E}h(W) - \mathbb{E}h(Z)|.$$

where Z is a standard Gaussian random variable. Let f_h be the solution to

$$f'(w) - wf(w) = h(w) - \mathbb{E}h(Z).$$

Replacing w by W and taking expectation on both sides of the above equation, we have

$$d_W(\mathcal{L}(W), N(0, 1)) = \sup_{\|h'\| \leq 1} |\mathbb{E}f'_h(W) - \mathbb{E}Wf(W)|. \quad (2.6)$$

If $\|h'\| \leq 1$, then it is known that (c.f. (2.14) of Raič (2004), Döbler (2012))

$$\|f'_h\| \leq \sqrt{\frac{2}{\pi}}, \quad \|f''_h\| \leq 1.$$

The bound (1.2) is proved by (2.6) and applying the above bounds in (2.4). \square

Remark 2.5. The following bound can be obtained following the proof of Theorem 1.1:

$$d_W(\mathcal{L}(W), N(0, 1)) \leq C_0 \left(\sqrt{\frac{c}{m}} + \frac{K_m}{\sqrt{cm^{3/2}}} + \frac{1}{m^{1/4}} \right)$$

where C_0 is an absolute constant, $K_m = \sum_{i=1}^m d_i$ and d_i is defined in Lemma 2.2. For fixed c , the above error bound goes to zero if $m \rightarrow \infty$ and $K_m \ll m^{3/2}$. This rules out complete graphs where $K_m \sim m^{3/2}$. Proposition 6.1 of Bhattacharya, Diaconis and Mukherjee (2013) gives this counter-example.

ACKNOWLEDGMENTS

We thank the authors of the paper Bhattacharya, Diaconis and Mukherjee (2013) for suggesting using Stein's method to give an alternative proof of their result. This work was supported by the NUS-Overseas Postdoctoral Fellowship from the National University of Singapore.

REFERENCES

- R. Arratia, L. Goldstein and L. Gordon (1990). Poisson approximation and the Chen-Stein method. With comments and a rejoinder by the authors. *Statist. Sci.* **5**, 403–434.
- A. D. Barbour and L. H. Y. Chen (2005). *An Introduction to Stein's method*. Lecture Notes Series **4**, Institute for Mathematical Sciences, National University of Singapore, Singapore University Press and World Scientific.
- A. D. Barbour, L. Holst and S. Janson (1992). *Poisson approximation*. Oxford Studies in Probability, **2**. Oxford Science Publications.
- B. Bhattacharya, P. Diaconis and S. Mukherjee (2013). Universal Poisson and normal limit theorems in graph coloring problems with connections to extremal combinatorics. *Preprint*. Available at <http://arxiv.org/abs/1310.2336>
- A. Cerquetti and S. Fortini (2006). A Poisson approximation for colored graphs under exchangeability. *Sankhya: The Indian Journal of Statistics* **68**, 183–197.
- S. Chatterjee, P. Diaconis and E. Meckes (2005). Exchangeable pairs and Poisson approximation. *Probab. Surv.* **2**, 64–106.
- L. H. Y. Chen (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3**, 535–545.
- C. Döbler (2012). Stein's method of exchangeable pairs for absolutely continuous, univariate distributions with applications to the Polya urn model. *Preprint*. Available at <http://arxiv.org/abs/1207.0533>
- X. Fang and A. Röllin (2014). Rates of convergence for multivariate normal approximation with applications to dense graphs and doubly indexed permutation statistics. *Bernoulli*. To appear.
- S. Janson and K. Nowicki (1991). The asymptotic distributions of generalized U -statistics with applications to random graphs. *Probab. Theory Related Fields* **90**, 341–375.
- M. Raič (2004). A multivariate CLT for decomposable random vectors with finite second moments. *J. Theoret. Probab.* **17**, 573–603.
- C. Stein (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Stat. Prob.* **2** Univ. California Press. Berkeley, Calif., 583–602.